Movie Release Overview – Killers of The Flower Moon
COMP 370 Final Project

Fall 2023 – December 11th, 2023
Karen Fu 260969951, Jessie Xu 260966881,
Livesh Armoogum 260967119, Raphael Julien 260966757
School of Computer Science
Faculty of Science
McGill University

Abstract

'Hollywood compressed theatrical performances into a packaged medium. And as a result, we no longer sling putrescent produce at stages. Nor do we usually clap after a good movie, not because we're indifferent to what we've seen but because there is no one there to applaud.'[1] says an unnamed author for The Atlantic. Now, Big Data allows data analysts to retrieve this lost feedback using public outlets.

Introduction

We are asked by a media coverage company to analyse the coverage and the reception of a movie, relative to other movies airing at a similar time. Considering the time frame of this assignment—that is November to December 2023—we chose the movie *Killers of the Flower Moon* (shortened to *KFM* for future reference), a blockbuster released on October 20th 2023, starring Leonardo DiCaprio among other high caliber hollywood stars. This movie had the benefit of being highly anticipated, which maximized our chances of gathering media posts about it, as well as being fresh in the mind of viewers and critics alike.

In order to measure and quantify the media coverage of a given article, we used the public API from NewsAPI.org[2], which allowed us to gather hundreds of news articles for free, and without whom this project would not have been possible.

The first step to our research was to refine the two tasks given to us by our stakeholder into quantifiable, computer-friendly statements that would guide our actions and decisions throughout the project. The final version of our refinements are as follow:

1. Given the topics we have selected, what label is the most relevant considering each article's title, description and content excerpt? (Done through Open Coding)
2. Given general query strings, how many "totalResults" does our movie's API call returns compared to other movies which aired on the same day? (October 20th 2023)

Data

Before explaining our data collection process, it is important to understand our phraseology. Throughout this report, we will refer to terms such as 'query string' and 'API call'. 'API call' refers to the content of the https GET request returned by the API (Figure 5), 'query string' refers to the contents of 'q=' (Figure 6). The Figures are found in the appendix.

Article Collection

Data for the 500 articles was collected using NewsAPI.org. Since the API returns articles that matches with the query string, the tricky part was to find the most adequate and precise set of keywords for our query string. Ideally, it should general enough such that all articles covering *KFM* are returned, but also discriminatory enough in order to filter out any unrelated article that may contain the set of keywords.

To ensure the preciseness of our API calls, we decided to combine the title of the movie with a set of keywords to are highly likely to be associated with it. Namely, we chose cast names, 'oscar', 'movie', 'premiere', 'theaters', and 'indigenous'. This wide variety of keywords ensured minimal bias, as many aspects of the movie would be covered by the articles returned. As such, our query string looked something like:

```
q={"Killers of the Flower Moon" AND
(Scorsese OR DiCaprio OR Gladhill OR
Fraser OR Movie OR Oscar OR Premiere OR
Theaters OR Indigenous)}.
```

In the string above, *KFM* is considered one uninterrupted string using the quotation marks. All spaces in the string were replaced with '%20' to match the https request's format. Lastly, many articles returned by the API appeared as 'Removed' or duplicates. Those articles were filtered out as to keep unique articles with content to extract.

Since this is a team project, it was paramount that no two teammates collected the same articles. We avoided this issue by splitting the article collection by mutually exclusive published times, where each member would focus on one time range. We split the published time as follow:

Member	From	To
Karen	Oct 24	Oct 31
Raphael	Nov 1	Nov 8
Livesh	Nov 9	Nov 16
Jessie	Nov 17	Nov 24

All those factors ensured a wide selection of articles likely related to our target movie, with minimal duplicates and parasite data.

Coverage Comparison

For the second part of this project, we selected all selected all movies that were released the same day as *KFM* listed on the website movieinsider.com. This gave us a list of 20 movies including ours that we could use for our comparison. We measured coverage using the *totalResult* field returned by the API call, which indicates how many articles matched the query string. To reduce bias, all API calls were done the same day (December 5th), and all query strings followed the same template:

- For movies whose title is unique (Multiple rare words, first names, neologisms, eg *Butcher's Crossing*) : `q={"<Movie Title>"}`
- For movies whose title is uncommon (reference to something else, common words, eg *Sick Girl*): `q={"<Movie Title>" AND Movie}`
- For movies whose title is common (common words, easily mistakable, title already exists, eg *The Post*): `q={"<Movie Title>" AND <CAST> AND Movie}`

This template allows for equal chance for any movie, regardless of their popularity, theme or cast.

Methods

Open Coding

We approached the design of our typology step as an iterative process. First, we sampled the first 50 articles of our respective dataset, and labeled according to the most prevalent theme of the article. We had a total of 21 distinct topics at first, which we removed, merged, and adjusted until we had our remaining eight. Below is a an example using 3 real articles articles from our dataset:

Title	Iter. 1	Iter. 2	Iter. 3
Every Leonardo DiCaprio Movie, Ranked	DiCaprio	DiCaprio	Actors
Robert De Niro tells jury that claims by ex-assistant are "nonsense"	Lawsuit DeNiro	DeNiro	Actors
Lily Gladstone in Gucci at the 2023 LACMA Art+Film Gala: IN or OUT?	Gladstone Festival	Gladstone	Actors

A significant design decision we took is that the articles were categorized regardless if they are about the movie *KFM* itself. In other words, we determined that articles on subjects corrolarry to the movie count as coverage, since the movie most likely influenced public interest in the first place. To give a concrete example, an article talking about Robert De Niro's lawsuit was labeled **Actors**: while the journal clearly is not about the movie, De Niro is a protagonist in it and the popularity of *KFM* undoubtedly ignited public interest in his personal life. Additionally, this prevents our dataset from having an overwhelming amount of unrelated articles.

Topic Characterization

We characterized our topics using the TF-IDF measure. This operation allows us calculate the frequency of a word present in a category, relative to the inverse frequency of the category. But why are we doing this? By applying the formula below, relevant words should rank high and inform our audience of what is meant by a topic name. We might also get surprises, which would indicate an inaccurate data annotation.

```
tf-idf = tf x idf
tf = frequency of word in topic
idf = log(# of topics/# of topics using w)
```

To avoid problematic behaviour, we chose to apply this formula only to the **title** and **description** of our articles, ignoring the **content** excerpt returned by the API. This is because the the content section was often infested with non-utf-8 characters and other escape sequences such as " that would make the processing harder and more tedious than necessary.

Coverage Comparison

We have selected several criteria to quantitatively compare the media coverage of the movies we selected:

- The totalResult field from the API call: this will naturally reveal whether or not the movie is being talked about or not
- The release medium (Theaters, Streaming, ...): this may indicate whether or not the movie's popularity was influenced by external factors such as advertisement, which differs from medium to medium
- The uniqueness of the title: this may indicate whether or not a movie's title can influence its availability on public outlets such as NewsAPI.org. Movies with unique titles are labeled **1**, common titles are labeled **3**.

Overall we recognized those factors to provide relevant insights into the online presence of the film *Killers of the Flower Moon*, as well as the efficiency of the various methods we used to conduct our research.

Results

Open Coding Topics

As mentioned in the **Methods** section, we tailored our categorization criteria such that 8 are left. We have selected the following typology in order to annote our dataset:

- Actors: Article focuses on some actor(s) of the movie

- Movie Review: Article/Blog post offering a reviewing/-giving its opinion on the movie
- Production: Article focuses on technical aspects of the movie
- Scorsese: Article focuses of the director Martin Scorsese
- Financial Performance: Article focuses on box-office or budget
- Themes Covered: Article focuses on plot, original book and themes
- Other movies: Article mentions *KFM* but is really about another movie
- Unrelated: Article is not about *KFM* at all

Note that we distinguished 'Other movies' and 'Unrelated' as we believed that an article mentioning *KFM* with respect to the movie industry still represented relevant coverage.

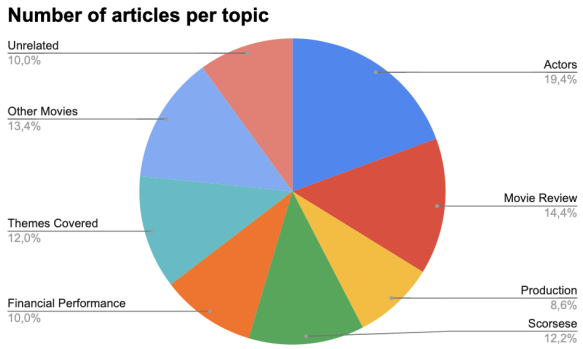

Figure 1: Distribution of topics over 500 articles

Most Used	Least Used	Mean	Std. Dev.
Actors (97)	Production (43)	12.5% (62.5)	3.16% (15.8)

Topic Characterization

For each category, we calculated the TF-IDF score of each word (neglecting stopwords). The 10 highest scoring words for each label is as follow:

Table 1: **Actors (Left)** and **Movie Review (Right)**

Rank	Word	Score	Word	Score
1	Assistant	106.05	Devery	68.62
2	Discrimination	68.62	Jacobs	68.62
3	Gender	45.75	Reservation	31.19
4	Former	39.23	Slams	22.87
5	De	36.19	Dogs	18.02
6	Brendan	35.31	Indigenous	18.02
7	Jury	33.27	Osage	16.64
8	Ex	31.19	Spielberg	12.48
9	Court	31.19	Review	11.09
10	Fraser	30.50	Palm	11.09

Table 2: **Production (Left)** and **Scorsese (Right)**

Rank	Word	Score	Word	Score
1	Intermissions	26.34	Joe	11.09
2	Sound	16.64	Russo	11.09
3	Intermission	15.24	Spielberg	11.09
4	Costume	12.48	Martin	11.08
5	Handful	10.40	Francesca	10.40
6	Theaters	8.93	Steven	9.70
7	Paramount	8.83	Scorsese	9.61
8	Screenings	8.32	Chanel	8.32
9	Inserting	8.32	Wolf	8.32
10	Tonight	8.32	John	8.32

Table 3: **Financial Performance (Left)** and **Themes Covered (Right)**

Rank	Word	Score	Word	Score
1	Box	34.31	Osage	30.50
2	Office	33.37	Native	22.87
3	Global	29.11	Indigenous	18.02
4	Swift	16.64	Murders	14.71
5	Eras	16.64	Warns	14.48
6	Taylor	13.86	Oil	11.09
7	Nights	13.73	Watching	11.09
8	Beats	12.47	Geoffrey	10.40
9	Million	12.22	Standing	10.40
10	Theatrical	9.70	Speak	10.40

Table 4: **Other Movies (Left)** and **Unrelated (Right)**

Rank	Word	Score	Word	Score
1	Napoleon	20.79	Strikes	9.70
2	Malkovich	20.79	Country	8.32
3	John	18.02	Fried	6.24
4	Box	17.39	Host	6.24
5	Office	17.39	Lachman	6.24
6	Ayo	16.63	Broke	6.24
7	Edebiri	16.63	Gift	6.24
8	Opus	16.63	Chalamet	5.88
9	a24	15.25	Quarter	5.54
10	Ridley	11.78	Deal	4.15

Coverage Comparison

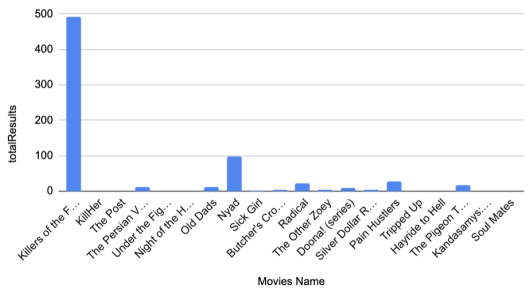

Figure 2: Media coverage per movie released Oct. 20th 2023

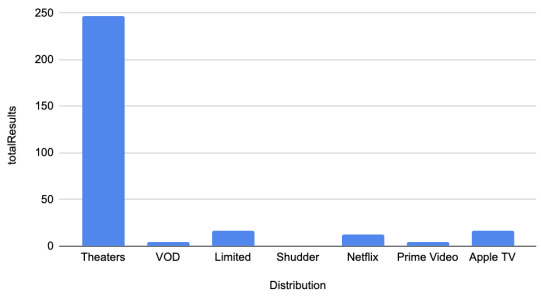

Figure 3: Average API result per release medium

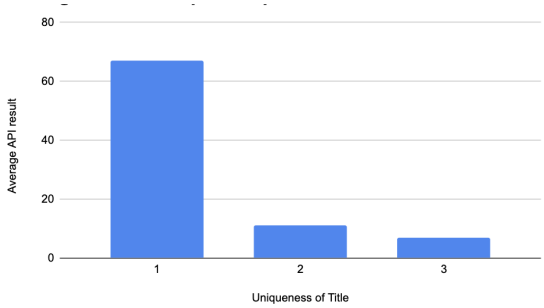

Figure 4: Average API result per uniqueness of title

Title	tR	Distrib.	Uniq
Killers of the Flower Moon	493	Theaters	1
KillHer	0	VOD	2
The Post	0	Theaters	3
The Persian Version	11	VOD	2
Under the Fig Trees	0	Limited	1
Night of the Hunted	0	Shudder	1
Old Dads	12	Netflix	3
Nyad	99	Limited	1
Sick Girl	1	VOD	2
Butcher's Crossing	3	Limited	1
Radical	23	Limited	3
The Other Zoey	4	Limited	1
Doona! (series)	9	Netflix	2
Silver Dollar Road	4	Prime Video	1
Pain Hustlers	28	Netflix	2
Tripped Up	0	Limited	3
Hayride to Hell	0	Limited	1
The Pigeon Tunnel	17	Apple TV	2
Kandasamys: The Baby	0	Netflix	1
Soul Mates	0	Limited	3

Table 5: Title, totalResult, Distribution, and Title uniqueness of all movies released on Oct. 20th 2023

Discussion

Open Coding

Firstly, it is important to notice that the distribution of our topics is relatively equal across all 500 data points (see Figure 1). This shows that all topics ended up being useful, and no topic was too niche or too general. Better yet, the 'Unrelated' topic ended up being the joint-second least used label with only 50 articles and below the mean, showing that it was not used as an 'other' or default section. It also shows the accuracy of our API calls, where 90% of our articles collected were related to our target subject.

Secondly, the **Actors** category leads the count with almost 20% of all articles belonging to it. This is not surprising considering the prevalence of celebrities in modern media as highly recognisable and marketable figures. With a cast consisting of Oscar-winning faces and other laureate recipients, *KFM* was bound to be receive high exposure. Another good decision was to split the director Scorsese on its own rather than merging it into a **Cast and Crew** category, something we considered.

Lastly, the **Production** category suffers the lowest count of articles with only 8.6% of the dataset. While this might be a proof of the modern public's fascination over celebrity drama over the technical aspects of movies, we believe that a this low number stems from the overlapping nature of the **Movie Review** and **Production** topics. Indeed, a review is likely to highlight technical qualities such as the Original Sound Track, the picture, the narrative, and the realizate; all of which could fall under **Production**. A solution could be to remove the **Movie Review** label and redistribute articles between the **Production** and **Themes Covered** sections.

Topic Characterization

The lines highlighted in bright yellow in Tables 1 through 4 are the keywords associated with the movie *Killer of the Flower Moon*. While many keywords seem to unrelated to the movie, it is important to remember our assumption that a keyword is categorized if it relates to its label. As such, many keywords indirectly related to the movie. Take the ACTORS category: while only "Brendan", "De [Niro]", and "Fraser" are actors in *KFM*, all other keywords are clearly tied with Robert De Niro's lawsuit with his ex assistant. By following this definition, we can recognize that each topic (besides **Unrelated**) possesses at least 4 related keywords and an average of 6,5. This is solid proof that the TF-IDF metric was successful in deciphering important statistics from our dataset and linking it directly with our typology.

Another positive observation is the fact that all top words relate to the topic, pointing to a successful annotation to our data. Here is an overview of each top word and how it related to their respective topic:

- (Actors) Assisstant: The word with highest score with a staggering 106,05. While it may seem surprising, it is easily explained by the prevalence of articles investigating Robert De Niro's lawsuit and the fact that this case is exclusive to the **Actors** category.
- (Movie Review) Devery: This refers to Devery Jacobs, an actress who vehemently criticized *KMF* on X.
- (Production) Intermissions: The lack of intermissions during projections in theater was a huge topic which en-

raged many on social media. This counts as production because it is a result of the movie's duration (over 3:30 hours) which is a production choice.

- (Scorsese) Joe: This refers to Joe Russo, a fellow director who notably directed Avengers Endgame, and whose loud feud with Martin Scorsese has instigated many reactions.
- (Financial Performance) Box: Box-Office performances are, have always been, and will forever be a hot topic of interest for report and movie consumers.
- (Themes Covered) Osage: Osage is the name of the indigenous tribe represented in the movie. This is perhaps the best term to illustrate the **Themes Covered** category.
- (Other Movies) Napoleon: Ridley Scott's *Napoleon* has featured in several posts along with *KMF*, mostly for comparisons or movie announcement posts.
- (Unrelated) Strikes: The hollywood writer's strike has paralysed the industry for several months, and *KMF* has not been spared. While we considered abstract the strikes in their own category, we realized it would not gather enough articles.

One flaw in our analysis is the management of the **Movie Review** category: besides having few keywords referring to a movie review, it seems to have many overlaps with the **Themes Covered** section (Osage and Indigenous). Perhaps the distinction between those two categories was unclear, and as such many articles were mis-classified. A better definition of the **Movie Review** category may have prevented such mis-classifications. Another theory is that many words associated with quality (good, better, bad, problem) were actually filtered out when removing stopwords. Laxing the stopword filter may improve results.

Coverage Comparison

Figure 2 Killers of the Flower Moon is by far the most covered movie. It itself overshadows all other films with 493 articles, against 211 combined. This difference is far from surprising: huge cast, big budget, highly anticipated. Nyad Pain Hustlers come second third with a large margin: still interesting and marketable cast (Judy Forster, Chris Evans) and Netflix release, boosting popularity. Other movies share little to no coverage whatsoever. There are two explanations for this phenomenon: either a lack of interest from the public local, lower-budget movies; or a lackluster coverage from the API itself.

Figure 3 Theaters win with a large margin, with an average of 248 articles from movies released in theaters, against a mere 50 combined. This is surprising considering the recent loss of popularity of the medium with the growth of streaming networks. That being said, this analysis is limited to 20 movies on 1-day: it would therefore be pertinent to extend such analysis to a larger scope.

Figure 4 We can see that the more unique the title, the more results the API will return. This shows that movies with extremely generic names such as "The Post" are more likely to be overlooked by the news API. This is because it is hard to target such movies using the query string. The API can easily miss media posts if the query string is too discriminatory, but it can also embed them with unrelated articles if the query string is too vague.

Conclusion

Throughout this case study, we have shown that Martin Scorsese's *Killers of the Flower Moon* has been in the **center of media attention** over movies released at the same time, with **over twice as much articles found**. By leveraging publicly available data processing tools, and by forming a solid and exhaustive typology, we have deduced that while various aspect of the movie are being discussed, **online journals focus their stories around cel and reviews of the picture**. Despite a somewhat lackluster API, we have succeeded in extracting significant information supporting our categorization using the TF-IDF method.

Group Members Contribution

All the data collection, data annotation, data analysis, and report redaction work was split equally among all 4 members in order to ensure equality of treatment. Each member contributed:

- The collection, formatting, filtering of 125 movie articles
- The annotation of 50, than all 125 movie articles according to a common typology
- The coverage analysis of 5 movies released on Oct. 20th
- The redaction of roughly 1.5 page of this report

Weekly meetings were held in order to report progress, request help, plan forward, and discuss the project. It is during those meetings that our respective data was merged into the main 500-articles dataset, and TF-IDF calculation was held. The code used at each step was unique to each member according to their preference, although code was shared upon request to speed up the process.

References

[1] Big Data and Hollywood: A love story. In: The Atlantic. https://www.theatlantic.com/sponsored/ibm-transformation-of-business/big-data-and-hollywood-a-love-story/277/. Accessed 27 Nov 2023

[2] NewsAPI.org documentation. https://newsapi.org/docs

Figure 5: Format of API response

```
'https://newsapi.org/v2/everything?q={}&from={}&to={}&language=en&apiKey={}
```

Figure 6: Format of http request sent to API